



SERVEI DE SISTEMES  
D'INFORMACIÓ GEOGRÀFICA  
I TELEDETECCIÓ  
Universitat de Girona

## IV JORNADAS DE SIG LIBRE

# Geolocalizador de noticias temáticas: el caso de los riesgos naturales.

J. Preciado López<sup>(1)</sup>, R. Berlanga Llavori<sup>(2)</sup>

<sup>(1)</sup> Consorcio Erasmus Mundus Master Geotech: IfGI-WWU-Münster, Alemania; Dept. Lenguajes y Sistemas Informáticos-UJI-Castellón, España; ISEGI-UNL-Portugal.  
[julioprecilopez@yahoo.com](mailto:julioprecilopez@yahoo.com)

<sup>(2)</sup> Departamento de Lenguajes y Sistemas Informáticos Universitat Jaume I, Campus de Riu Sec E-12071 Castelló de la Plana, España. [berlanga@lsi.uji.es](mailto:berlanga@lsi.uji.es)

## RESUMEN

*Las bases de datos geoespaciales temáticas en distintas escalas geográficas y temporales, son necesarias en multitud de líneas de investigación. Una de ellas es la gestión y alerta temprana de riesgos de desastres por amenazas naturales (inundaciones, huracanes, terremotos, etc.). Las noticias sobre éste tema se publican habitualmente en periódicos digitales de todo el mundo y comportan un alto contenido geográfico. Este trabajo pretende extraer automáticamente las noticias emitidas por canales de re-difusión web (conocidos por las siglas RSS en inglés) para georreferenciarlas, almacenarlas y distribuir las como datos geoespaciales. Mediante técnicas de procesamiento de lenguaje natural y consultas a bases de datos de topónimos realizaremos la extracción de la información. El caso de estudio se aplicará para México y todos los componentes utilizados serán de código abierto.*

**Palabras clave:** Bases de datos geoespaciales, noticias, RSS, lenguaje natural, extracción, código abierto.

## ABSTRACT

*The thematic geospatial databases in different geographic and temporal scales are needed in many areas of research. One is risk and disaster management. The stories on this subject are published regularly online by newspapers worldwide, entailing a high geographical content. This work aims to automatically extract the news releases from web feeds (RSS/Atom) to stored and distributed them as geospatial data. For that we use natural language processing techniques. The use case applies to Mexico and all the components used are open source.*

**Key words:** Geospatial databases, news, RSS, natural language processing, open source.

## INTRODUCCIÓN

La creación de bases de datos espaciales es costosa en términos de tiempo y recursos materiales y económicos. Los geodatos son una parte fundamental de las Infraestructuras de datos espaciales (IDEs). En este trabajo presentamos la extracción de geodatos referentes al dominio de los riesgos naturales.

En esa línea, el riesgo se define como el grado de combinación entre amenazas y vulnerabilidades [1]. Las primeras se refieren a los fenómenos naturales que pueden causar daño en diferentes lugares, escalas de tiempo y grados de intensidad [2]. Amenazas o peligros naturales comunes son: terremotos, movimientos de terreno (deslizamientos, caídas, etc.), huracanes, fuertes lluvias, sequías e inundaciones entre otros. Por su parte, la vulnerabilidad básicamente se expresa por la fragilidad de las comunidades, sus economías e infraestructuras físicas ante aquellos peligros [3]. Entonces un desastre natural es una serie ruptura activada por algún fenómeno natural peligroso (por ejemplo un huracán) que causa serios daños medioambientales y pérdidas materiales y humanas, excediendo la capacidad de respuesta de los afectados [1].

En numerosos países no existen registros para estudiar el riesgo o su acceso público es restringido. Muchas veces los datos que hay están a escalas muy generales (nivel regional estatal). Esa carencia de datos impide elaborar índices de riesgo detallados que afectan la utilidad de los resultados [4]. Además de la escala geográfica, los eventos que se registran suelen ser aquellos que en cuestión de horas provocan grandes pérdidas materiales y humanas, dejando a un lado los eventos medianos y pequeños que en suma podrían ser más costosos que los primeros.

Basado en lo anterior nuestra propuesta es extraer información de la web para crear bases de datos espaciales. Las cuales podrán distribuirse mediante geoservicios y ser explotadas por la comunidad científica interesada en los riesgos naturales.

Los objetivos propuestos son: 1) extraer información textual a partir de periódicos digitales (Online); 2) almacenar esa información de forma estructurada; y 3) hacerla accesible para cualquier geo-cliente.

Los siguientes puntos introducen algunos conceptos y trabajos relacionados a la recuperación y extracción de información. Posteriormente explicamos la metodología que incluye la arquitectura, el modelo de datos y el análisis de las noticias. Finalmente explicamos los resultados y ofrecemos una conclusión.

## CONCEPTOS Y TRABAJO PREVIO

Este trabajo está estrechamente vinculada a la recuperación y extracción de información. Usa técnicas de minería de datos, específicamente de minería de texto. Esta última se ayuda de técnicas de procesamiento de lenguaje natural (PLN). Una de ellas es el reconocimiento de entidades nombradas (NER por sus siglas en inglés). La técnica NER nos ayuda a localizar e identificar entidades en textos desestructurados. Hay varios enfoques NER, el que utilizamos parte de los sistemas basados en conocimiento.

Para enlazar las técnicas de minería de texto y el procesamiento de datos geográficos es necesario recordar algunos conceptos como geoparsing, geocoding y geotagging. El primero se refiere a procesar cualquier archivo con un contenido

geográfico [5]. El segundo trata de codificar geográficamente un objeto a partir de otras entidades geográficas como códigos postales, calles y direcciones [6]. El tercero trata de identificar y anotar contenido geográfico a ficheros sin referencia espacial, tales como fotografías, textos y vídeos [7].

Otros términos acuñados por la comunidad extracción y recuperación de información geográfica son Geo y No-geo ambigüedad. No-geo ambigüedad sucede cuando los nombres de lugares son los mismos que las palabras comunes [8] Por ejemplo León como provincia de España y como el apellido de una persona. Geo-ambigüedad ocurre cuando el mismo termino coincide con más de un lugar [9]. Por ejemplo, "Las Vegas" es una ciudad muy famosa en los Estados Unidos pero aparece más de cinco veces en los resultados después de una búsqueda en Geonames [10].

Existen innumerables trabajos relacionados con el tema de recuperación y extracción de información. De los primeros podemos citar algunos que consideramos destacados. Por ejemplo, de las contribuciones pioneras que involucran recuperación de información geográfica están [11]; en cuanto a desambigüedad de topónimos se puede ver [9]. En la misma línea pero utilizando métodos de aprendizaje automático supervisado para la Geo y No-geo ambigüedad ver [12]. En cuanto al geotagging de fotografías podemos citar [13] , [14] y [15].

En cuanto a los trabajos enfocados a la extracción de información se encuentra [16] entre los primeros. Por su parte [17] ha experimentado extraer nombres vernáculos de sitios web sociales para aumentar las entradas de topónimos en gaceteros tradicionales.

Recientemente hay numerosos trabajos que utilizan Web feeds para recuperar y extraer información. Para geodatos se utilizan específicamente GeoFeeds. Los GeoFeed o GeoRSS son documentos RSS o Atom anotados geográficamente. Existen dos formatos, uno simple y otro extendido y perfil del GML (Geography Markup Language) [18]. Tal vez el primer sitio que publicó GeoFeeds fue [19]. Otros ejemplos son [20], dedicado a referenciar noticias del Reino Unido mediante un mapa. Otra importante contribución es Geonames [21]. Éste proyecto incluye un amplio catalogo de millones de nombres de lugares habitados y rasgos geográficos libres para descargar. La comunidad de Geonames ha desarrollado varios servicios, entre ellos destaca el convertidor RSS/Atom a GeoFeed.

Como se puede ver, hay una gran diversidad de proyectos y métodos para recuperar y extraer información de la web. A continuación presentamos el nuestro.

## **METODOLOGÍA**

La metodología esta dividida en varias secciones. En principio se describe brevemente la arquitectura del sistema. Posteriormente describimos el modelo de datos y el caso de uso. Finalmente analizamos el método de extracción de la información y los resultados obtenidos a la fecha.

### **Arquitectura**

La arquitectura está diseñada bajo el paradigma cliente-servidor. Nuestro sistema comprende básicamente cuatro componentes (Figura 1). La primera consiste en un agente que lee los feeds de los periódicos digitales. Ellos se utilizan como insumo para ser analizados en el segundo componente. Éste a su vez utiliza los catálogos para filtrar (remover la Geo y No-geo ambigüedad, clasificar y anotar) los feeds.

Con esos resultados se puebla la base de datos espacial. Esta a su vez es accesible por medio de un servicio OGC<sup>1</sup>.

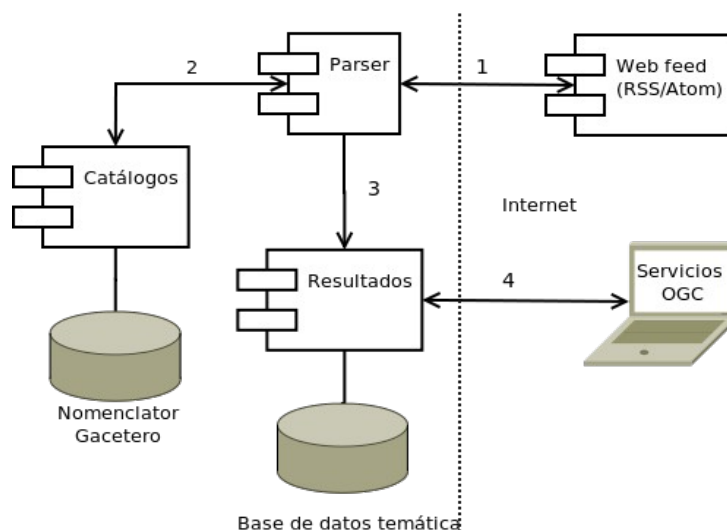


Figura 1: Arquitectura del sistema

## Modelo de datos

El modelo de datos está diseñado sobre la premisa del caso de uso descrito en la introducción. Como prototipo se prevé probarlo con la división administrativa de la República Mexicana. El modelo está dividido en tres paquetes, cada uno con diferentes clases. El primer paquete, por su origen llamado Marco Geoestadístico Nacional (MGN) incluye la clase Lugar de la cual depende la clase Estado. Los estados se componen al menos de un municipio. De forma similar, los municipios se componen de localidades. Para el caso de México, esta división jerárquica (estado/municipio/localidad) está codificada por el INEGI (Instituto Nacional de Estadística, Geografía e Informática [22]) como sigue: dos dígitos para cada una de las 32 entidades estatales (01:32); tres para los municipios (001:999) y cuatro para las localidades (0001:9999). La Tabla 1 y la Figura 2 muestran esa jerarquización con el estado de Tamaulipas como ejemplo. Actualmente México cuenta con 32 estados, 2456 municipios y más de 200 mil localidades [22].

Tabla 1: Codificación de entidades según el MGN.

Nivel administrativo.	Clave		Nombre	Rango
Estado	00	28	Tamaulipas	28:28
Municipio	000	005	Burgos	28001: 28043
Localidad	0000	0052	Las Flores	280010001: 280430510
EE+MMM+LLLLL	00000000	280050052	Tamaulipas/ Burgos/ Las Flores	No aplica

A pesar de que los estado y municipios son polígonos, en la base de datos final sólo se registraran sus centroides.

El otro paquete siguiente es llamado Marco de Riesgo. El cual contiene las clases Evento y Desastre, ambas descritas en las Tablas 3 y 2.

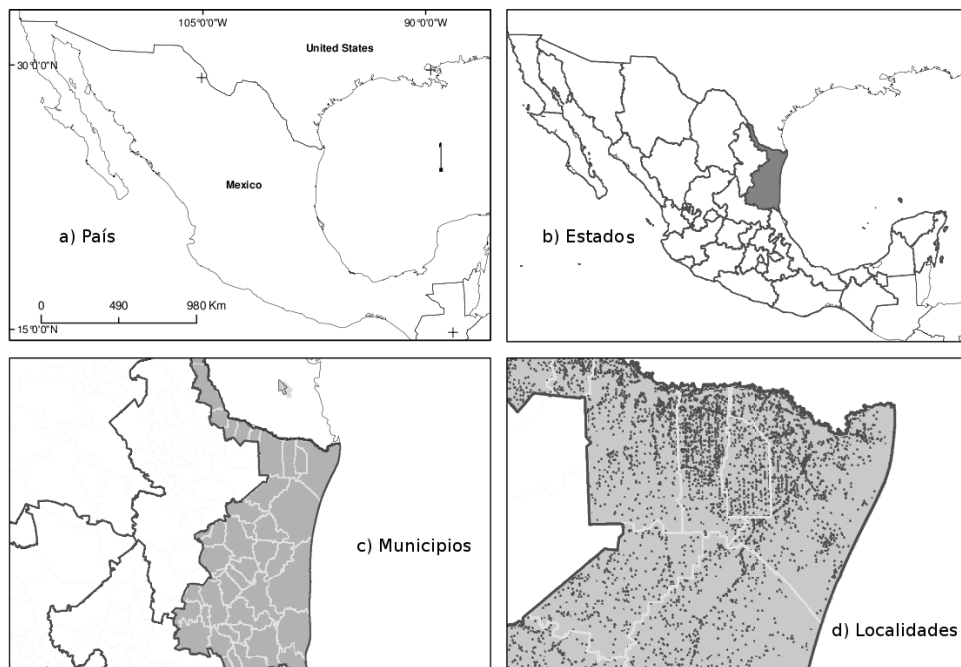


Figura 2: División administrativa de México

Finalmente se incluye un paquete complementario nombrado Marco Temático por el carácter de información que contiene. En él se incluyen capas que pueden ayudar a contextualizar visualmente los puntos donde ocurrieron los eventos y desastres. Las capas que se incluyen de diferentes fuentes son: población, índice de marginación, suelos, geología, uso de suelo, altitud e inclinación del relieve.

### Análisis de noticias

El método que se va a emplear para analizar las noticias se realiza en dos partes. La primera incluye un procesador para desambiguar los topónimos y reconocer entidades de texto relacionadas con amenazas naturales. Cuando este procesador no es del todo exitoso, se pasa a otro para analizar las dependencias gramaticales en los textos. Dado que nuestro método es extraer información relevante a partir de los textos dentro de las noticias publicadas como web feeds, iniciamos el análisis explicando las características fundamentales de dichos formatos.

### Web feeds

Los Web feeds se utilizan para proporcionar alertas cuando un sitio web se ha modificado. Se trata de una manera simple de leer y escribir información en la web. Por lo tanto, un usuario no tiene necesidad de comprobar las actualizaciones, sino un agente emite boletines alertando si un sitio ha sido alterado. Los Web feeds son documentos XML<sup>2</sup> y heredan todas sus características (seguridad, estructura, expansibilidad, etc.). Existen dos formatos de Web feeds: RSS (Really Simply Syndication) y Atom. Ambos los hemos venido llamando indistintamente como Web feeds o feeds.

---

### 2 Extensible Markup Language

Plaça Ferrater Mora 1, 17071 Girona

Tel. 972 41 80 39, Fax. 972 41 82 30

[infojornadas@sigte.udg.es](mailto:infojornadas@sigte.udg.es) <http://www.sigte.udg.es/jornadassiglibre/>

A pesar de que el formato Atom es más completo y "mejor" estructurado, el RSS por ser más antiguo es más difundido y utilizado. Ambos comparten una estructura semejante. Aquí revisamos brevemente la anatomía de un documento RSS versión 2.0; a su vez es la noticia que utilizaremos como ejemplo para el análisis.

El fichero tiene un elemento al nivel superior denominado "RSS" como sí mismo (negritas Figura 3). Dentro del elemento RSS se anida otro único llamado "Channel" (negritas y cursivas). Dentro de "Channel" se listan elementos adicionales, de los cuales título ("title"), vínculo ("link") y descripción ("description") son requeridos. Dentro de "Channel" otro elemento opcional es "item" en él se describe cada historia correspondiente a una noticia. El elemento "item" a su vez incluye otros como: título, link fecha de publicación y descripción (cursivas).

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
    <title>La Jornada</title>
    <pubDate>Wed, 17 Sep 2008 09:05:42 GMT</pubDate>
    <description>Noticias del diario mexicano La Jornada</description>
    <link>http://www.jornada.unam.mx/2009/09/17/</link>
  <item>
    <title>Provoca lke inundaciones y tres muertes en Nuevo León</title>
    <link>http://www.jornada.unam.mx/2008/09/17/index.phpsection=estados
&article=039n1est&partner=rss</link>
    <pubDate>Wed, 17 Sep 2008 09:05:42 GMT</pubDate>
    <description>Las lluvias que provocó el huracán lke causaron la muerte
de tres personas en la sierra del municipio de Santiago, Nuevo León, y desbordaron la
presa El Cuchillo, lo que inundó durante más de 24 horas algunos tramos de las
carreteras que van de Monterrey a Reynosa y a Ciudad Victoria, y de ésta a San Luis
Potosí. En Reynosa, el oleaje arrasó con al menos 100 metros de la barda de contención
de playa Bagdad y devastó el área de palapas.</description>
  </item>
</channel>
</rss>
```

Figura 3: Documento RSS

En el ejemplo se presenta una historia. La mayoría de los periódicos digitales incluyen sólo un extracto o el primer párrafo de la historia completa en el elemento "description".

A partir de los elementos del feed se pueden extraer los nombres de los lugares donde ocurrió un evento, la fecha y sus atributos. Para identificar los atributos cabe destacar el estilo de redacción que utilizan periodistas para documentar las noticias. Por ejemplo, casi siempre escriben en mayúsculas los nombres propios y normalmente describen lugares siguiendo las divisiones administrativas (por ejemplo: Girona, Cataluña, España).

#### Primer proceso

El proceso de la extracción de noticias se hizo con un corpus de prueba con cien noticias del mismo periódico digital de la Figura 3 (fuente [23]). Las historias corresponden a diferentes fechas. La mitad de ellas menciona algún evento como lluvias e inundaciones que causaron algún daño a las poblaciones. La otra mitad son irrelevantes.



El procesamiento se divide en dos fases, la primera es la utilización de una herramienta que le llamamos Geo-identificador. Su objetivo es reconocer, identificar y clasificar entidades de nuestro interés en el texto. Para ello se apoya en heurística y métodos basados expresiones regulares, listas "stopwords", el gacetero del MGN y otras listas predefinidas que incluyen componentes del riesgo como: lluvias intensas, tormentas tropicales, huracanes, terremotos y otros relacionados con aspectos sociales y estructurales, tales como: pobreza, asentamientos inadecuados, lugares e infraestructuras inseguras. Asimismo, se proporcionan nombres de organizaciones de ayuda ante desastres (Cruz Roja) y otras agencias del gobierno (Protección Civil). El Geo-identificador básicamente nos ayuda a eliminar la Geo-ambigüedad, clasificar y registrar las entidades de texto que nos interesan. Para ello resuelve los casos de ambigüedad a través del nomenclátor del MGN. Después, aísla los topónimos y asigna los atributos a cada uno de esos nombres de lugares. Las restricciones para el caso es que exista al menos un lugar de nuestro gacetero en las noticias y varios términos relacionados al contexto de los riesgos y desastres naturales. Una vez cargado el feed en memoria, el parser realiza los siguientes pasos:

1. El procesador recorre el título y la descripción de la historia para determinar si hay topónimos en el texto.
2. Comparar los nombres en el catálogo de MGN.
3. Si se encuentran topónimos, almacena los candidatos y sus rangos de clave, de lo contrario termina.
4. Después de identificar los topónimos, el programa empieza a verificar administrativamente de arriba hacia abajo (estado a localidad) si hay más nombres de lugares en las noticias dentro del rango de claves precedente.
5. Si se encuentra topónimos a un nivel administrativo inferior, va al paso siguiente, de lo contrario termina.
6. La iteración anterior se repite en los municipios para detectar localidades.

El Geo-identificador analiza los bloques de texto separados por puntos como entidades distintas. Consecuentemente de la misma historia se pueden generar varios registros de lugares en diferentes niveles administrativos.

El Geo-identificador proporciona dos resultados importantes. El primero es una salida para generar tuplas y llenar por vez primera la base de datos. La segunda son bloques de texto anotados cuando no fue posible identificar cuál de los múltiples lugares corresponden los atributos, es decir una relación muchos (topónimos) a muchos (atributos). En tal caso los bloques etiquetados se analizan en el siguiente paso.

### **Segundo proceso**

El análisis de dependencias gramaticales (ADG) lo utilizamos para determinar las relaciones de palabras a través de las oraciones. Para el ADG utilizamos un procesador de dependencias incluido en "Freeling". La cual es una biblioteca de código abierto que proporciona servicios para análisis computacional del lenguaje [24] Esta herramienta maneja alrededor de siete idiomas incluyendo el Español. Se apoya en las reglas gramaticales de cada idioma y otros diccionarios de conocimiento.

Las entidades anotadas por Geo-identificador las procesamos con el ADG de Freeling. Después del análisis gramatical, el resultado devuelto es un árbol de dependencias.

El ejemplo de la Figura 4 es el árbol del texto debajo (titulo del feed):

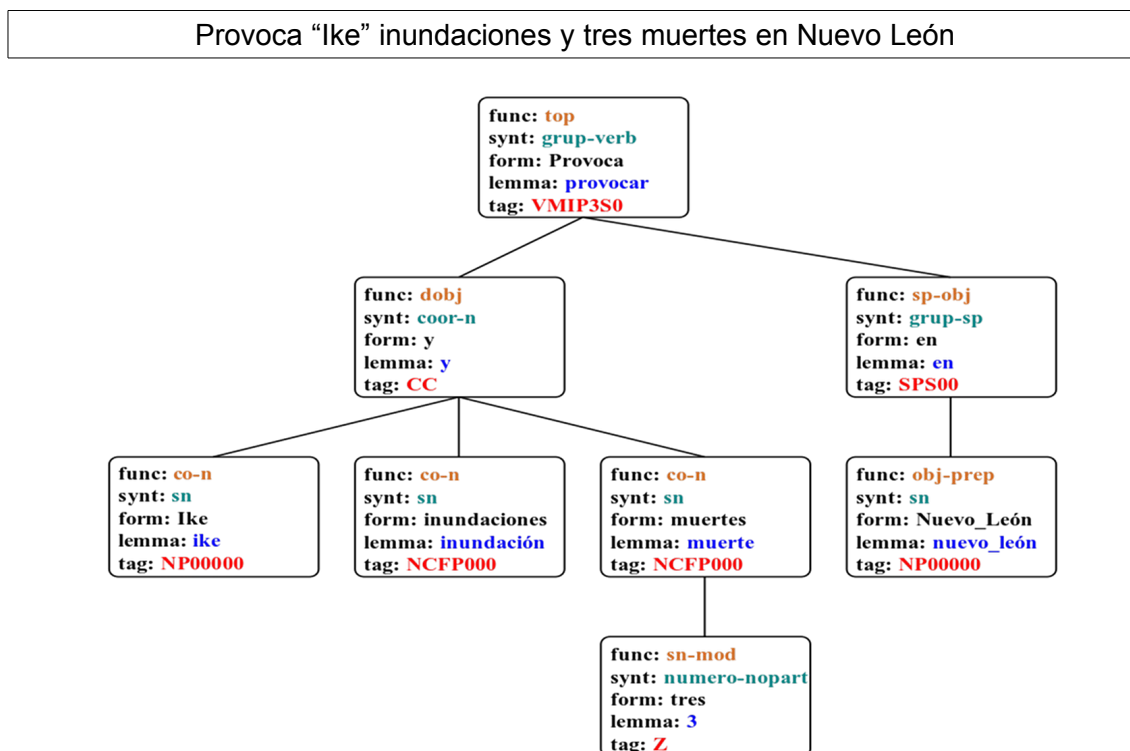


Figura 4: Árbol de dependencias

El árbol se parsea y se procesa para crear las tuplas que se omitieron en el Geo-identificador. Con ellas se puebla la base de datos (Tablas 2 y 3) para georreferenciar los datos Figura 5.

Tabla 2: Tabla Desastres

Tabla Desastre		
Columna	Descripción	Ejemplo
causa	Causa del evento e impacto	Lluvias, huracán
impacto	Objetos dañados o destruidos (i.e., cientos de hectáreas de cultivos perdidos, barrios incomunicados, playas arrasadas, etc.).	
daño	Millones de US\$/€	
víctima	Numero de desplazados, damnificados, muertos o personas afectadas.	muerte de tres personas
ayuda	Organismo que otorgo alguna ayuda (por ejemplo: Naciones Unidas).	



Tabla 3: Eventos

Tabla Evento		
Columna	Descripción	Ejemplo
cveMgn	Clave del lugar según el Marco Geostadístico Nacional (MGN)	19049
nomMgn	Nombre del lugar según MGN	Santiago
rasgoMgn	Nivel administrativo en el MGN	Municipio
lat	Latitud	25.5166
lon	Longitud	-99.6666
IDeve	Id. del evento = cveMgn+fecha	1904917192008
fecha	DD/MM/AAAA	17192008
nomEve	Nombre del fenómeno natural	inundaciones
titulo	Titulo de la noticia en el feed	Provoca Ike inundaciones y tres muertes en Nuevo León
desc	Descripción de la noticia en el feed	Las lluvias que provocó el huracán Ike causaron la muerte de tres personas en la sierra del municipio de Santiago, Nuevo León, y desbordaron la presa El Cuchillo, lo que inundó durante más de 24 horas algunos tramos de las carreteras que van de Monterrey a Reynosa y a Ciudad Victoria, y de ésta a San Luis Potosí. En Reynosa, el oleaje arrasó con al menos 100 metros de la barda de contención de playa Bagdad y devastó el área de palapas.
lugarRel	Otros lugares relacionados en la noticia	Sierra del municipio de Santiago, El Cuchillo, Ciudad Victoria, San Luis Potosí
rasgoRel	Rasgos (ríos, lagos, sierras) relacionados en la noticia	Presa, carreteras, Sierra.
nomPropio	Nombres propios encontrados	Ike
fuelle	Link a la noticia	<a href="http://www.jornada.unam.mx/2008/09/17/index.phpsection=estados&amp;article=039n1est&amp;partner=rss">http://www.jornada.unam.mx/2008/09/17/index.phpsection=estados&amp;article=039n1est&amp;partner=rss</a>

## RESULTADOS

Los primeros resultados obtenidos arrojan al rededor de un 80% de éxito considerando el corpus de noticias. Sin embargo, al modificar algunos parámetros en el Geo-identificador se ha mejorado la precisión en las tuplas pero a la fecha no se ha estimado por completo su grado de confianza.

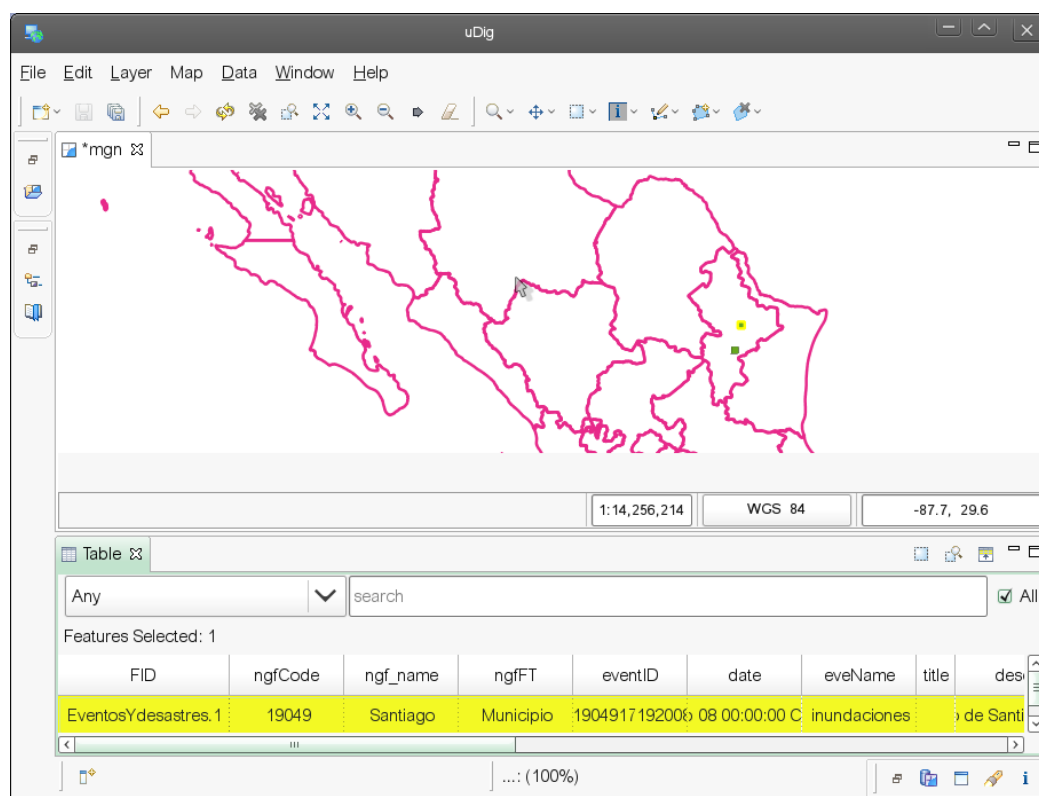


Figura 5: Noticia georreferenciada

## CONCLUSIONES

Siguiendo el problema de la carencia de geodatos a distintas escalas temporales y geográficas, nos adjuntamos a la tarea de explorar métodos que permitan generar esa información a partir de la Web. Para ello utilizamos Web feeds de los periódicos digitales, procesándolos bajo las consideraciones de un dominio específico. Los resultados han sido un tanto satisfactorios. Sin embargo, hay mucho trabajo futuro por realizar. La conexión entre geodatos e información textual se podría mejorar implementando ontologías y semánticas en los gaceteros de nombres y rasgos geográficos. De manera similar, incluyendo procesos multidimensionales utilizando la tecnología OLAP.

## AGRADECIMIENTOS

Al comité de coordinadores Erasmus Mundus Geotech por la beca para realizar el Master. A Rafael Berlanga por su gran atención, cooperación y disponibilidad de enseñanza. A todos mis compañeros y amigos con los que he discutido y compartido mis ideas.

## REFERENCIAS

[1] UNDP (2004) Reducing Disaster Risk a Challenge for Development, United Nations Development Programme, Bureau for Crisis Prevention and Recovery One United Nations Plaza, New York, NY 10017, USA.

[2] Blaikie P, Cannon T, Davis I, Wisner B. At Risk: Natural Hazards, People's Vulnerability, and Disasters. London, UK: Routledge, 1994

[3] Inter-American Development Bank. Indicators of Disaster Risk and Risk Management Program for Latin America and The Caribbean. Summary report. World Conference on Disaster Reduction. Kobe, Hyogo, Japan, 2005.

[4] Margaret Arnold , Robert S. Chen , Uwe Deichmann , Maxx Dilley , Arthur L. Lerner-Lam , Natural Disaster Hotspots Case Studies Details. The World Bank Hazard Management Unit. Washington, D.C. 2006

[5] Geoparsing - Wikipedia, the free encyclopedia. Available at: <http://en.wikipedia.org/wiki/Geoparsing> [Accessed November 29, 2009].

[6] Geocoding - Wikipedia, the free encyclopedia. Available at: <http://en.wikipedia.org/wiki/Geocoding> [Accessed November 29, 2009].

[7] Geotagging - Wikipedia, the free encyclopedia. Available at: <http://en.wikipedia.org/wiki/Geotagging> [Accessed December 1, 2009].

[8] Smith, David A. (2002) Detecting and Browsing Events in Unstructured text, SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 73–80

[9] Smith, David A., and Crane, Gregory (2001) Disambiguating Geographic Names in a Historical Digital Library, ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, Springer-Verlag, 127–136

[10] GeoNames. Available at: <http://www.geonames.org/> [Accessed November 16, 2009].

[11] Woodruff Allison G., and Plaunt Christian (1994) GIPSY: Automated geographic indexing of text documents. Journal of the American Society for Information Science, Vol. 45. No 9, Pp, 645-655.

[12] Hu, You-Heng, and Ge, Linlin (2007). A Supervised Machine Learning Approach to Toponym Disambiguation, The Geospatial Web, 117–128

[13] Popescu, Adrian, and Grefenstette, Gregory (2009): Deducing trip related information from flickr, WWW '09: Proceedings of the 18th international conference on World wide web, ACM, 1183–1184

[14] Flickr: Explore everyone's photos on a Map. Available at: <http://www.flickr.com/map/> [Accessed November 11, 2009].

[15] Ahern, Shane and Naaman, Mor and Nair, Rahul and Yang, Jeannie Hui-I (2007). World explorer: visualizing aggregate data from unstructured text in geo-referenced collections, JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ACM, 1–10.

[16] Sundheim, Beth M. (1995). Overview of results of the MUC-6 evaluation, MUC6 '95: Proceedings of the 6th conference on Message understanding, Association for Computational Linguistics, 13–31.

[17] Twaroch, Florian A., Smart, Philip D., and Jones, Christopher B. (2008). Mining the web to detect place names, GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval, ACM, 43–44.

[18] GeoRSS. Available at: <http://www.georss.org/> [Accessed November 16, 2009].

[19] ACME Available at: <http://www.acme.com> [Accessed December 14, 2009].

[20] BBC News feed for UK in Google Maps Available at: <http://dev.benedictoneill.com/bbc> [Accessed January 7, 2010].

[21] Wick, Marc, and Becker, Torsten: Enhancing RSS Feeds with Extracted Geospatial Information for Further Processing and Visualization, The Geospatial Web, 105–115, 2007

[22] Instituto Nacional de Estadística y Geografía (INEGI). Available at: <http://www.inegi.org.mx/inegi/default.aspx> [Accessed October 28, 2009].

[23] <http://www.jornada.unam.mx>

[24] Jordi Carrera, Irene Castellón Marina Lloberes Lluís Padró, and Tinkova, Nevena (2008). Dependency Grammars in FreeLing, Procesamiento del Lenguaje Natural, 21–28.